単語連鎖列書き換え規則を用いた形態素解析後処理とその評価

佐々木尚人　　宮崎 正弘

新潟大学大学院自然科学研究科

1　研究の背景

形態素解析システム Maja [1] は単語を短単位の組み合わせに分割する。例えば「によって」という格助詞相当語は、格助詞「に」動詞「よ」活用語尾「っ」助動詞「て」という短単位の分割になっている。ここで音声処理に対しては短単位の分割が望ましく機械翻訳に対しては長単位の分割、先ほどの「によって」の例ならひとまとめに格助詞とする分割が望ましい。このことから Majaで短単位に分割された単語を用途によって使い分けるための形態素解析後処理が必要と考えられる。

2　形態素解析後処理の位置づけ

図１に形態素解析の位置づけを示す。
入力された日本語文は Maja により形態素に分割される。次にそれを形態素解析後処理に渡し、単語連鎖列書き換え規則により、用途に見合った単語連鎖列に書き換えたものを構文解析に渡す。

3　日本語形態素解析システム Maja の流れ

形態素解析後処理の説明の前に Maja の流れを説明する。

例えば「私は学生です。」という日本語文を形態素解析システムに入力すると、まず一般語辞書から辞書引きする、辞書検索、数詞の生成が行われる。次に接続ルールを使い、先ほど辞書引きされた品詞が文法的に接続できるかチェックする接続処理が行われる。文末まで解析が成功しなかった場合には、固有名詞辞書再検索、未知語解析を行い、文末まで解析を成功させる。 未知語解析では部分的再試行機構により未知語の抽出と品詞推定を行う。 最後に複合名詞解析を行ない結果を出力する。 同形語や単語分割の曖昧性はコスト最小法により絞り込む。

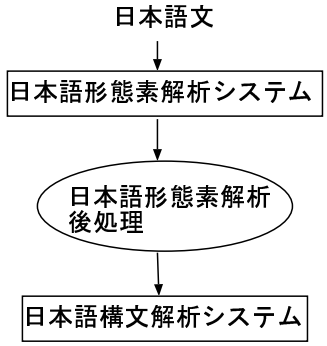

図 1: 形態素解析後処理の位置づけ

4　コスト最小法による曖昧性の絞り込み

それぞれの形態素間の接続の強さは異なる。動詞の語幹–語尾などは接続が強く、 名詞–副詞などの接続は弱いと考えられる。このような接続の強さを接続コストとして数値で与え、 その接続コストを用いて曖昧性を絞り込み、 優先順位を

Postprocessing for Japanese Morphological Analysis using Word String Rewriting Rules
Naoto Sasaki , Masahiro Miyazaki
Niigata University

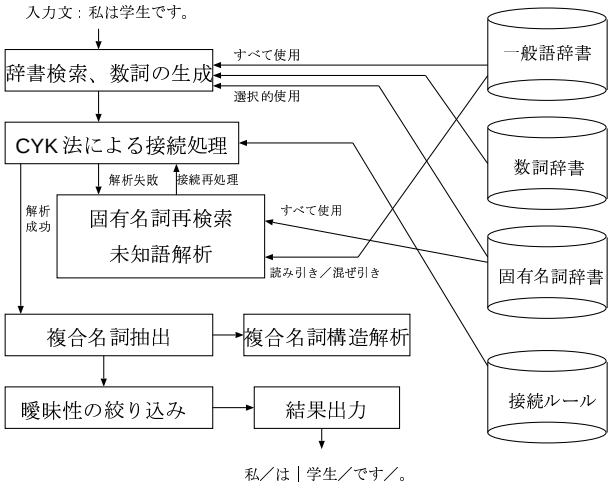

図 2: 形態素解析システムの流れ

つける。この方法はグラフ構造の各ノードまでの最小の部分コスト及びそのパスを記憶しておく。この処理を文末まで行い、文末のノードに記憶されているパスがコスト最小パスである。

5 Majaの性質と後処理の目的

日本語形態素解析 Maja は高速性、汎用性を基にしているので、接続処理は隣り合った 2 単語間の関係で行っている。また形態素の分割は短単位で行っている。

そのため、機械翻訳に適した形態素解析にするには、離れた単語間の関係、3 単語以上連鎖での関係を考慮する必要がある。また長単位の分割にする必要がある。

これらを形態素解析後処理の目的とする。

6 後処理が必要な例

後処理が必要になる文は以下のようないくつかの単語によって格助詞や助動詞相当語が生成できるものである。

- それ/に/よ/っ/て/起こ/っ/た/。
  によって → 格助詞相当語
- そこ/に/入/る/こと/が/でき/た/。
  ことができ → 助動詞相当語
- 走/る/と/とても/息/が/乱れ/る/。
  と → 格助詞 Or 接続助詞
  乱れ →「と」と離れた位置にある条件になりうる品詞

7 単語連鎖書き換え規則の形式

単語連鎖書き換え規則の形式を以下に示す。

$W_{\ i}$(pos,sem,sup),*,---,*,$W_{\ 1}$(pos,sem,sup),*,
[W_1(pos,sem,sup),---,W_j(pos,sem,sup)]
,*,W_{+1}(pos,sem,sup),*,---,*,W_{+k}(pos,sem,sup)
→
$W_{\ i}$(pos,sem,sup),*,---,*,$W_{\ 1}$(pos,sem,sup),*,
$W_{1'}$(pos,sem,sup),---,$W_{j'}$(pos,sem,sup)
,*,W_{+1}(pos,sem,sup),*,---,*,W_{+k}(pos,sem,sup)
{ } .

ここで、W_m(pos,sem,sup) は字面 W_m の単語の品詞が pos、意味属性が sem、補足条件が sup であることを表す。pos は必須の引数、sem、sup は省略可能な引数である。W_m(pos,sem,sup) の代わりに ⟨W_{m1}(pos,sem,sup)/---/W_{mn}(pos,sem,sup)⟩ と記述することにより、n 個の同形語 W_{m1},---,W_{mn} の OR を表す。[ ] 内は書き換え対象となる単語連鎖列である。[ ] の前後にある単語連鎖列は書き換え対象単語連鎖列の前後に共起する単語連鎖列を規定するもので、省略可能である。＊は、この部分に長さ 0 以上の任意の単語連鎖列が位置することを示している。{ } は補強項であり、必要に応じて本規則の適用条件を記述する。

本規則により、長さ 1 以上の単語連鎖列 W_1,---,W_j の前方の長さ 1 以上の単語連鎖列 (構成要素が不連続の離散型でも可) が $W_{\ i}$,---,$W_{\ 1}$ で、後方の長さ 1 以上の単語連鎖列 (離散型でも可) が W_{+1},---,W_{+k} で、補強項の条件を満たす場合、単語連鎖列 W_1,---,W_j を $W_{1'}$,---,$W_{j'}$ に書き換える。

8　単語書き換え規則の適用

単語書き換え規則の形式に従い、

- 連語の抽出
  W_1,---,$W_j \rightarrow W_{1'}$
- 同形語の判別
  W_{m1}/---/$W_{mn} \rightarrow W_{m'}$

を行い用途に見合った単語連鎖列に書き換える。

9　連語の抽出の例

単語書き換え規則において、長さ２以上の単語連鎖列 W_1,---,W_j を一つの連語 $W_{1'}$ に書き換えることによって、以下の例に示すように、短単位語に分割された単語連鎖列を一つの連語にまとめることができる。

- 格助詞相当語：
  [に (格助詞), よ (五段ラ行動詞),っ(活用語尾), て (既定の助動詞)]
  → によって (格助詞,< 仕手/仲介/手段/根拠/原因 >).
- 複合辞：
  [こと (形式名詞), が (格助詞), でき＊ (一般動詞)]
  → ことができ＊ (助動詞, 可能).
  （＊は活用語尾を表すワイルドカード文字）
- それ以外の連語：
  [こと (普通名詞), に (格助詞), よ (五段ラ行動詞), る (活用語尾), と (接続助詞)]
  → ことによると (副詞, 推量).

10　同形語の判別の例

単語書き換え規則を用いて、長さ１以上の単語連鎖列 W_1,---,W_j 中の単語の同形語セット $\langle W_{m1}/\text{---}/W_{mn}\rangle$ を一つの単語 $W_{m'}$ に絞り込むように書き換えることによって、以下の例に示すように、同形語の前後に共起する単語連鎖列 (連続型 / 離散型) によって同形語を判別することができる。

「鳩は動物**で**、鳥である。」のように、「で」の直後が読点で、後方に名詞＋断定の助動詞がある場合、断定の助動詞と判別。
W_{1}(名詞), [⟨で (格助詞)/で (断定の助動詞)⟩], W_{+1}(読点), *, W_{+2}(名詞), W_{+3}(断定の助動詞)
→ W_{1}(名詞), で（断定の助動詞）,W_{+1}(読点),*,W_{+2}(名詞),W_{+3}(断定の助動詞)
「その事故は偶然によるもの**で**故意によるものでない」のように「で」の直前の名詞と同じ名詞＋断定の助動詞が後方にある場合、断定の助動詞と判別。
W_{1}(名詞), [⟨で (格助詞)/で (断定の助動詞)⟩], *, W_{+1}(名詞), W_{+2}(断定の助動詞)
→W_{1}(名詞), で (断定の助動詞),*,W_{+1}(名詞),W_{+2}(断定の助動詞) $\{W\_{1} = W_{+1}\}$.

11　形態素誤りの修正の例

単語書き換え規則を用いて、以下の例のような特定の単語連鎖において生じる単語認定や単語区切りの誤りを修正できる。

- 品詞認定の誤り修正
  [で (格助詞), も (係助詞), あ＊ (五段ラ行動詞)]
  →[で (断定の助動詞), も (係助詞), あ＊ (断定の助動詞)].
- 単語区切りの誤り修正
  [残念 (形動名詞), な (助動詞), がら (普通名詞)]
  → [残念 (形動名詞), ながら (接続助詞)]

12　同形語判別の適用結果

同形語の判別に対しての評価実験を行った結果を表１に示す。

表１：同形語判別の評価実験結果

適用条件	適用数	正解数
	85	81
A	70	70
B	3	0
C	2	2
D	8	8
E	2	1

・ A
＊ (名詞) ＋で (格助詞) ＋は (係助詞) ＋な (形容詞) ＊ (活用語尾) →
＊ (名詞) ＋で (助動詞) ＋は (係助詞) ＋な (形容詞) ＊ (活用語尾)

・ B
＊ (活用語の終止形) ＋と (格助詞) ＋*＋＊（動詞、と格をとらない）→
＊ (活用語の終止形) ＋と (接続助詞) ＋*＋＊（動詞、と格をとらない）

・ C
＊ (名詞) ＋とも (副助詞) ＋*＋＊ (動詞、と格とる) →
＊ (名詞) ＋と (格助詞) も (係助詞) ＋*＋＊ (動詞、と格とる)

・ D
＊ (名詞) ＋でも (係助詞) ＋あ (動詞) ＋＊ (活用語尾) →
＊ (名詞) ＋で (助動詞) も (係助詞) ＋あ＊ (助動詞

・ E
＊１ (名詞) ＋で (格助詞) ＋*＋＊２ (名詞) ＋＊ (断定の助動詞) →
＊１ (名詞), で (断定の助動詞) ＋*＋＊２ (名詞) ＋＊ (断定の助動詞) {＊１＝＊２} .

13　他の形態素解析システムの出力への変換例

単語書き換え規則を用いて、 Maja の出力結果を、品詞体系や辞書への単語収録単位の異なる他の形態素解析システムの出力に変換できる。

・ Maja 出力の ALTJAWS 出力への変換例
その本は読むことはできない。
その/本/は/読む/こと/は/でき/ない
→ その/本/は/読む/ことはでき/ない

14　まとめ

形態素解析後処理を行い、単語書き換え規則を適用して以下のことを行った。

・ 連語の抽出

・ 同形語の判別

・ 形態素誤りの修正

・ 他の形態素解析システムの出力への変換

15　今後の課題

・ 単語書き換えの条件の充実

・ 単語情報の付与

・ 書き換え規則の適用範囲の再考

参考文献

[1] 尾嶋、宮崎：日本語形態素解析システムにおける部分的再試行機構の導入とその効果、情報処理学会第 58 回全国大会 1E-4(1999)

[2] 白井、横尾、松尾、大山：日英翻訳のための日本語解析技術、NTT R&D、Vol.48、No.12、pp.1399–1404(1997)